Investigating the dimensions of spelling ability. [IF: 1.661]

Notenboom, A.; Reitsma, P.

published in
Education and Psychological Measurement
2003

DOI (link to publisher)
10.1177/0013164403258442

document version
Publisher's PDF, also known as Version of record

Educational and Psychological Measurement
http://epm.sagepub.com/

Investigating the Dimensions of Spelling Ability

Annelise Notenboom and Pieter Reitsma

Educational and Psychological Measurement 2003 63: 1039
DOI: 10.1177/0013164403258442

INVESTIGATING THE DIMENSIONS OF SPELLING ABILITY

ANNELISE NOTENBOOM AND PIETER REITSMA
PI Research—Vrije Universiteit Amsterdam

In this study, the latent structure of a spelling achievement test for elementary school grades was investigated. Factor analyses revealed that for Grades 2 to 6, the test was unidimensional, whereas for Grade 1, two factors were found: a phonological factor and an orthographic factor. Because of the anchor-test design, vertical equating was required to fit a two-parameter, unidimensional item response theory model. Concurrent calibration of the parameters, implemented in BILOG–Multiple Groups (MG), was used as the linking procedure. An evaluation of the model showed that invariant parameter estimates and ability scores were obtained. The transition from an alphabetic to an orthographic strategy occurs in the first grade. Learning to spell throughout elementary education mainly consists of learning to apply orthographic knowledge.

Keywords: *spelling ability; item response theory; vertical equating; spelling test*

Since a few decades, learning to spell has been considered to be a developmental process that involves the integration of several skills (Templeton & Morris, 2000). Phonemic awareness and the knowledge of grapheme-phoneme correspondences are critical in the early stages of learning to spell. Later, grammatical and morphological knowledge, analogies with words in lexical memory, and the knowledge of orthographic rules and conventions become important as well (Caravolas, Hulme, & Snowling, 2001; Lennox & Siegel, 1994).

Several developmental models for spelling have been presented (Frith, 1980, 1985; Henderson & Templeton, 1986; Marsh, Friedman, Welch, &

This research is supported by a grant from the Dutch Organization for Scientific Research to Pieter Reitsma (No. NWO 411-21-006). The authors gratefully acknowledge the helpful suggestions on item response theory modeling provided by Dr. Dato de Gruijter. Correspondence concerning this article should be addressed to Annelise Notenboom or Pieter Reitsma, PI Research, P.O. Box 366, 1115 ZH Duivendrecht, the Netherlands; e-mail: a.notenboom@psy.vu.nl or p.reitsma@psy.vu.nl.

Educational and Psychological Measurement, Vol. 63 No. 6, December 2003 1039-1059
DOI: 10.1177/0013164403258442

Desberg, 1980), mainly based on research in the English-speaking community. In these models, spelling is conceived as progressing through a series of qualitatively distinct stages in which different sources of knowledge are used. There is less agreement among these models about how to describe and characterize the development over time and how the different stages are to be qualified. Research on spelling indicates that various skills and sources of information are primarily mediated by two different processes (Lennox & Siegel, 1994). One is a phonological process in which spelling is based on the application of the phoneme-grapheme correspondence rules, and the other is based on a process in which lexical analogies can be facilitated through a process of direct lexical access. The developmental shift from the stage of phonological spelling to a spelling based on lexical knowledge is supposed to occur around the fifth grade (Marsh et al., 1980). Then, pupils start using an analogy strategy in the spelling of unknown words.

However, the stage-like models are not without controversies. The stages do not appear distinct in practice, and the distinction might be too simplistic. Therefore, some researchers abandoned the stage-like characterization of spelling development (Varnhagen, McCallum, & Burstow, 1997). The variety of strategies children use in spelling may probably be better described in terms of overlapping waves (Siegler, 2000). Over time, different strategies are developed that differ in frequency of use. Although at a particular moment, one strategy may predominate, it is not the only one available. The development of spelling ability is a continuous process, reflecting gradual improvements in children's phonological and orthographic knowledge. From the beginning, children use all the information available for the spelling of words. What may appear to characterize a particular stage is in reality a preponderance of use of one or the other strategy rather than use of one strategy in an absolute sense. For example, Treiman (1994) found that first graders already honor the orthographic patterns of English, even when they are not explicitly taught at school. Thus, children begin learning about the orthographic structure of their language at an early age. Varnhagen et al. (1997) showed that the change from phonological to correct spelling is gradual and does not show an obvious shift from one strategy to another.

Moreover, Brown and Loosemore's (1994) model of spelling development leaves the idea of the use of different strategies altogether. According to this model, which is based on research in connectionism, the process of learning to spell can usefully be viewed as one of mastering a set of statistical associations between representations of phonological forms of words and representations of their orthographies. This connectionist model assumes that just one process underlies the spelling of both phonologically regular and phonologically irregular words.

In this study, we conceptualize the issue of spelling development in a different way while taking advantage of models derived from test theory. Most research on the development of spelling is performed in a (quasi-)experimental fashion in which the items presented to participants are carefully chosen such that they represent spelling categories—phonological or orthographic characteristics—that are of concern to the experimenter (see, e.g., Marsh et al., 1980; Varnhagen et al., 1997). In this study, we used a psychometric approach. Based on psychometric theory, it is possible to conceive the constructs measured by a test as latent variables. A standardized spelling test is the focus of this study. The main question is concerned about the latent structure of this test. Can spelling ability be conceived as one single latent variable, or do the variety of skills related to phonology, orthography, the application of rules, and morphology that are required to spell a word correctly form different subdimensions? Furthermore, are the two main stages related to phonological and lexical knowledge discernable as two distinct, latent subcomponents of the general trait of spelling ability? And if they are, is it possible to mark a point on the developmental continuum, for example, following Marsh et al. (1980) at the fifth grade, where a shift occurs from the preponderance of use of the phonological to the lexical strategy?

This study deals with the spelling of Dutch. The question is whether the theories of spelling development, which are based primarily on research in the English-speaking community, are applicable to other languages. Like English, the spelling of Dutch is mainly based on the alphabetic principle, although it is more regular. Instruction in the Netherlands typically involves the spelling of phonologically regular words. Approximately from the end of the first grade onward, pupils are introduced to orthographic patterns that deviate from phonological transparency, the application of spelling rules, the role of silent letters, and inconsistent phoneme-grapheme mappings.

A test that covers all facets of spelling categories designed for all the elementary grades is the basis for this study. Factor analysis and item response theory (IRT) modeling (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968) were used to address the dimensionality of its latent structure.

Developmental Score Scales and Vertical Equating

The PI-dictee (Geelhoed & Reitsma, 1999), the spelling test that is the focus of this study, measures spelling achievement throughout the elementary grades, thus defining a developmental continuum for spelling ability. The test items gradually increase in difficulty for the grades. Because of this developmental continuum, it is not feasible to administer all items to all pupils. Therefore, we presented a subset of items to the pupils tailored to specific grade levels. A presentation of the items that are presented to the different grades is given in Table 1. However, administering items tailored to the level of ability requires a statistical procedure of bringing test scores back to the same scale. The process of determining the relationship between scores on two or more test forms is commonly referred to as equating (Skaggs & Lissitz, 1986). There are two types of equating: horizontal equating and vertical equating. Vertical equating is applicable for test forms that are intentionally different in difficulty, and that are administrated to different examinee populations of varying ability levels.

Vertical equating is not without controversies. How well are we able to measure human growth? Different subsets of items might constitute different skills, and the question is whether they can reasonably be scaled to be on a single dimension (Skaggs & Lissitz, 1986). Apart from such theoretical concerns, there are studies that focus on the evaluation of methods that can be used in vertical equating. Relevant here are those that explicitly modeled (sub)tests of varying degrees of difficulty that were administered to groups of unequal ability levels within the context of IRT modeling. The studies reported mixed results. Holmes (1982) and Loyd and Hoover (1980) found vertical equating to be inadequate for the one-parameter model; Forsyth, Saisangjan, and Gilmer (1981), on the other hand, reported satisfactory results for the one-parameter model, although assumptions, including unidimensionality and a common level of discrimination of the items, were not met.

Several other studies compared equating within the IRT context to traditional methods. In a study by Harris (1991), IRT true score equating did not provide better results than did equipercentile methods. Kolen (1981) compared true score and observed score IRT equating for both the one-, two-, and three-parameter model to traditional equipercentile linear methods for two tests. No method turned out to be superior, although in general, the equating based on IRT models yielded better results. Furthermore, Guskey (1981) found traditional grade-equivalent scale estimates to be less adequate than is IRT-based equating in a one-parameter model.

What all these studies have in common is that first, separate models were estimated for different groups, and then the parameter estimates were linked by means of an equating method. More recently, a computer program, BILOG–Multiple Groups (MG) (Zimowski, Muraki, Mislevy, & Bock, 1996), has been developed that implements concurrent calibration of the item parameters. Separate prior distributions on the ability vectors are allowed for the different grades during the item calibration phase. Then, the estimates of the item parameters are generated so they are on a common scale, making further linking unnecessary. A formal exposure can be found in Bock and Zimowski (1995) and Mislevy (1987).

Kim and Cohen (1998) compared concurrent calibration methods with the approach based on separate calibrations using equating coefficients from the

Table 1
The Schedule of Item Administration

Block	5	10	15	20	25	30	40	50	60
Item Numbers	1 to 15	16 to 30	31 to 45	46 to 60	61 to 75	76 to 90	91 to 105	106 to 120	120 to 135
Grade 1	X	X	X						
Grade 2	X	X	X	X	X				
Grade 3			X	X	X	X	X		
Grade 4				X	X	X	X	X	
Grade 5					X	X	X	X	X
Grade 6					X	X	X	X	X

characteristic curve method. They used a simulated data set consisting of two groups differing in ability level. They found that in cases of a larger number of common items, concurrent calibration methods yielded similar results as did the characteristic curve method. However, in cases of a small number of common items, linking using the characteristic curve method outperformed the concurrent calibrations.

Using simulated data that violated the assumption of unidimensionality, Béguin and Hanson (2001) evaluated separate and concurrent calibration for vertical equating. They found that in general, the concurrent estimation method resulted in a smaller bias. However, they did not systematically vary the number of common items (20% of the total test in their study).

Because of the complexity of our data set (e.g., several items were administered to more than two groups of different ability levels), we used the concurrent calibration methods implemented in BILOG-MG.

Research Questions

In this study, we will investigate the dimensionality of spelling ability. Does spelling ability consist of just one latent trait, or is it composed of more interpretable subdimensions? Is it possible to find a developmental shift, marking the preponderance of one strategy over another? Related to this issue is the question of whether the IRT modeling, taking departure from the idea of latent traits, forms a feasible approach. Furthermore, the issue of vertical equating is addressed in this study. Is it possible to bring the subsets of items, administered to groups of different ability, to the same scale? How well are we able to construct a developmental continuum of spelling ability throughout the elementary grades?

Method

Materials

The test that is the focus of this study is called PI-dictee (Geelhoed & Reitsma, 1999) and is a well-known and widely used measure for assessing spelling ability in the Netherlands. The PI-dictee is a standardized spelling test consisting of parallel forms (A and B), each containing 135 words that gradually increase in difficulty. Both monosyllabic and polysyllabic words are included that are commonly known to children of their age.

Following the standard administration procedure, the words are dictated to the pupil within a sentence; only the target word has to be written down. The words are grouped into nine blocks indicating varying levels of spelling ability. For example, the words of the first block are all orthographically transparent and can be spelled correctly by the average pupil who has received 5 months of education, whereas the words of the ninth block are indicative of the average spelling ability of a pupil who has received 50 months of education. Starting from the second block, words are included that show some deviance from phonological transparency, such as words that contain specific orthographic patterns or silent letters or words that require the application of some rules or lexical knowledge. Table 2 is a summary of the PI-dictee.

Participants

Participants were 3,657 pupils from Grades 1 to 6 of 22 elementary schools in the Netherlands. The schools were representative on background variables such as social class and geographical region. The sample of pupils included 1,775 girls and 1,705 boys, with 177 pupils having missing data on sex. Boys and girls were approximately equally divided across grades. No data on race or social class of the individual pupils were available.

Procedure

Not all items of the PI-dictee were presented to all pupils. The selection was dependent on grade and was tailored to the ability level. Table 1 presents the subsets of items administered to the grade levels. Because both forms (A and B) were used, twice as many items as mentioned in Table 1 were actually administered. The subsets of items were chosen in such a way that the range of ability of the pupils of a particular grade was mostly covered. This procedure deviates from the standard administration procedure, which requires that the test should be continued unless eight or more mistakes within one

Table 2
Characterization of the PI-Dictee

	Block								
	5	10	15	20	25	30	40	50	60
Phonological transparency	X	X	X	X	X	X	X	X	X
Orthographic patterns		X	X	X	X	X	X	X	X
Application of spelling rules			X	X	X	X	X	X	X
Inconsistent phoneme-grapheme mappings			X	X	X	X	X	X	X
Loan words							X	X	X

block have been made. However, for practical and efficiency reasons, this procedure was not followed. Consequently, for highly superior or inferior achieving pupils, the measurements might be less reliable.

The design of data collection used here is called the anchor-test design (Petersen, Kolen, & Hoover, 1989). Although the administration of items is tailored to grade levels, adjacent grades have some items (anchor items) in common. These anchor items should satisfy some requirements for equating to be adequate. They need to represent the content and statistical characteristics of the total test, and within the educational context, their number should be at least 20% of the total test length (Kolen & Brennan, 1995). In a study by Cook, Eignor, and Taft (1988), the results obtained from the vertical equating process varied considerably depending on which common items were used. From Table 1, it is clear that all items have been presented to at least two adjacent grades, suggesting that the equating procedure implemented here should be fairly stable. The teachers administered the tests in one or two sessions. Three independent raters marked the items. The interrater reliability was high (.98).

Statistical Analyses

To investigate the dimensionality of spelling ability, for each grade, an exploratory factor analysis was performed on the matrix of tetrachoric correlations of the items. Furthermore, for each grade, both the one- and two-parameter IRT models were estimated with BILOG (Mislevy & Bock, 1982). Most of the default settings of this computer program were used. Omitted items were scored as wrong answers. A normal distribution of spelling ability was assumed for each grade. Because the PI-dictee consists of free response items, the guessing parameter was assumed to be zero.

Based on the parameter estimates of the best fitting IRT models, for each grade, a data set that was known to be unidimensional was created. Given the parameter estimates, the probability of answering an item correctly can be computed; the data were simulated in such a way that an item was scored correctly if this probability was higher than a random number between 0 and 1. The simulated data sets had the same number of pupils and items as did the real ones. Furthermore, factor analyses on the tetrachoric correlations of the simulated item pairs were performed. In line with the modified parallel analysis proposed by Drasgow and Lissak (1983), the eigenvalues of the factor analyses on the real and simulated data sets were compared; multidimensionality is indicated when the second eigenvalue of the real data is substantially larger than the second eigenvalue of the simulated data.

All IRT models that involved vertical equating were estimated with BILOG-MG (Zimowski et al., 1996). The pupils of the different grades were treated as sampled from different populations. Simultaneous scaling was used as equating procedure, and the item parameters of the subtests were estimated jointly in one analysis. The item parameters were estimated by means of the marginal maximum likelihood method. In vertical equating over a range of age levels, the ability distributions of the groups may be widely spaced; therefore, a large number of quadrature points (30) was used. The ability distributions were estimated in the form of a discrete distribution on the quadrature points. Because the origins and unit of the ability distribution are up to linear transformations, they have to be fixed in the calibration either by setting the mean and standard deviation of a reference group to 0 and 1, respectively, or by setting the mean and standard deviation of the combined groups. In this study, Grade 1 served as the reference group in the calibration of the items.

The ability parameter estimates were obtained by means of the Bayesian expected a posteriori estimator. The program derived prior information regarding the latent distribution to which examinees belonged during an earlier estimation phase. The convergence criterion for the expectation maximization iterations (used as method to solve the marginal maximum likelihood solution) was set to .001. Again, omitted items were scored as wrong answers.

Results

Separate IRT Models for Each Grade

For each grade, the one-parameter and two-parameter models with one latent trait were estimated. The fit indices for the models are given in Table 3, and they strongly suggest that the two-parameter model was more appropriate for all grades.

Table 3
Fit Indices

Grade	Negative Log Likelihood One-Parameter Model	Negative Log Likelihood Two-Parameter Model	Difference	*df*	*p* Value
1	49,639.6	48,775.0	864.6	90	.000
2	73,896.2	73,039.7	856.5	150	.000
3	68,305.7	67,570.2	735.5	150	.000
4	63,012.5	62,209.5	803.0	150	.000
5	71,084.7	70,013.3	1,071.4	150	.000
6	59,873.2	58,671.3	1,201.9	150	.000
All	396,359.3	387,531.1	8,828.2	270	.000

Unidimensionality and Factor Analysis

Because of the better fit of the two-parameter model for all grades, data were simulated by means of the discrimination, difficulty, and ability parameters obtained from these models. Then, we computed tetrachoric correlations and eigenvalues from both the simulated and real data sets for each per grade. Scree plots of the 10 largest eigenvalues from both the simulated and real data sets appear in Figure 1. For Grade 2 to Grade 6, examination of the plots revealed that the PI-dictee was highly dominated by one latent factor, which accounted for 40% to 50% of the variance. Obviously, the second factor forms the breaking point in the scree plots. It is important that the factor analyses on the simulated data sets of Grade 2 to Grade 6 showed highly similar results.

For Grade 1, the situation was quite different. The first factor of the real data was considerably smaller than the one obtained from the simulated data, and there was a large discrepancy in value for the second eigenvalues. Therefore, for Grade 1, the conclusion seemed to be justified that the data set was two dimensional.

After a PROMAX rotation of the two-factor solution for Grade 1, the elements of the pattern and structure matrices were inspected. Items with pattern and/or structure loadings equal to or higher than .60 on one of the factors are presented in Table 4. The correlation between the two factors was .27.

Items loading on the first factor were mainly phonologically transparent, whereas the items loading on the second factor showed some deviation from phonological spelling. For example, the *-d* in /vriend/ (friend), pronounced as *-t*, is spelled as such because of the morphological relation to the plural form /vrienden/, and the patterns *-oei* and *-aai* can be phonologically correct spelled as *-oej* and *-aaj*; the grapheme *e* serving the role as a schwa is phonologically irregular. All items with such deviations appeared to be strongly connected to the second factor.

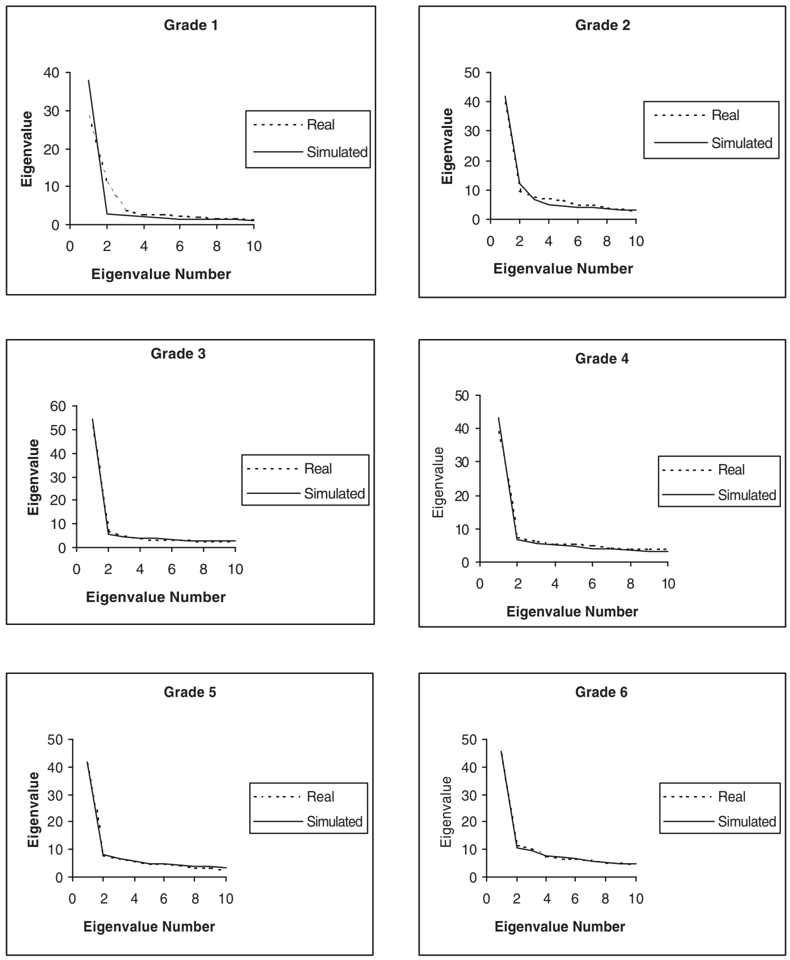

Figure 1. Factor solutions of the real and simulated data sets for Grades 1 to 6.

Remember that the 90 items completed by the first graders form a subsample of the 150 items that were completed by the second graders; therefore, it was quite remarkable that the solutions of these two grades did not agree. An additional analysis based on the smaller number of 90 items was performed for the second graders. This analysis did not show a dominant second factor, and the rotated pattern loadings did not reflect a difference in items that are related to phonology and orthography. So, there was no clear interpretation of the second factor of the 90 items completed by the second graders.

Table 4
Oblique Primary Factors for Grade 1

		Pattern		Structure		
Item	Translation	Factor I	Factor II	Factor I	Factor II	Nonphonological Spelling Pattern
In	In	.19	.80[a]	.41	.85[a]	
Duur	Expensive	.64[a]	–.01	.63[a]	.16	
Mos	Moss	.68[a]	–.20	.62[a]	–.02	
Pijl	Arrow	.67[a]	–.16	.62[a]	.01	
Fiets	Bike	.52	.30	.60[a]	.44	
Hert	Deer	.73[a]	.03	.74[a]	.24	
Zwart	Black	.59	.20	.65[a]	.36	
Op	On	.95[a]	–.56	.79[a]	–.30	
Doos	Box	.81[a]	–.39	.69[a]	–.17	
Week	Week	.82[a]	–.24	.75[a]	–.01	
Zuur	Sour	.68[a]	–.14	.63[a]	.03	
Heg	Hedge	.79[a]	–.19	.74[a]	.02	
Jeuk	Itch	.67[a]	–.17	.62[a]	.01	
Rups	Caterpillar	.69[a]	.08	.71[a]	.28	
Koets	Coach	.64[a]	.26	.72[a]	.44	
Slim	Clever	.89[a]	–.05	.88[a]	.19	
Brief	Letter	.67[a]	.08	.70[a]	.28	
Start	Tail	.73[a]	–.02	.72[a]	.18	
Plant	Plant	.76[a]	.02	.77[a]	.23	
Schuur	Shed	.66[a]	.12	.69[a]	.30	-sch
Fruit	Fruit	.71[a]	–.08	.69[a]	.10	
Slok	Swallow	.92[a]	–.19	.87[a]	.06	
Schaats	Skate	.54	.36	.64[a]	.51	-sch
Meer	Lake	.53	.33	.62[a]	.47	
Klapdeur	Self-closing door	.59	.25	.62[a]	.37	
Spoor	Track	.56	.22	.67[a]	.42	
Fietsbel	Bike bell	.58	.23	.65[a]	.39	
Schaal	Dish	.60[a]	.15	.64[a]	.32	-sch
Weer	Weather	.63[a]	.28	.71[a]	.46	
Slot	Locker	.85[a]	–.09	.83[a]	.14	
Door	Through	.59	.22	.65[a]	.39	
Paard	Horse	.12	.63[a]	.30	.67[a]	-d
Bloei	Bloom	.02	.74[a]	.23	.76[a]	-oei
Sprong	Jump	.33	.56	.49	.65[a]	
Beste	Best	.25	.59	.41	.66[a]	-d
Zwaard	Sword	–.06	.73[a]	.14	.71[a]	-d
Kinderen	Children	.22	.67[a]	.40	.72[a]	schwa
Strand	Beach	–.10	.81[a]	.12	.78[a]	-d
Draai	Turn	–.06	.72[a]	.13	.70[a]	-aai
Werkster	Cleaning lady	.11	.66[a]	.29	.69[a]	schwa
Busje	Little bus	.23	.62[a]	.40	.68[a]	schwa
Hoofd	Head	–.25	.83[a]	–.02	.76[a]	-d

(continued)

Table 4: (continued)

Item	Translation	Pattern Factor I	Pattern Factor II	Structure Factor I	Structure Factor II	Nonphonological Spelling Pattern
Zwaai	Swing	–.17	.79[a]	.05	.74[a]	-aai
Ruimte	Space	.05	.64[a]	.23	.66[a]	schwa
Vriend	Friend	–.28	.85[a]	–.05	.77[a]	-d
Gisteren	Yesterday	.08	.69[a]	.28	.71[a]	schwa
Sprookje	Fairy tale	.11	.76[a]	.32	.79[a]	schwa
Groei	Growth	–.18	.78[a]	.03	.72[a]	-oei
Zusje	Sister	.07	.71[a]	.27	.73[a]	schwa
Mond	Mouth	–.17	.70[a]	.02	.65[a]	-d

a. Pattern and/or structure loadings that are equal to or greater than .60.

Assessing the Fit of the IRT Model

While making use of the vertical equating procedures in BILOG-MG, we fitted one model for all grades. The program converged after 313 iterations (using a criterion of .001). Fit indices of the one- and two-parameter IRT models are given in Table 3. Analogous to the separate analyses for the grades, the two-parameter model provided a better fit to the data.

When an item response model fits to the data, invariant examinee ability estimates and item parameter estimates are obtained. The first feature of the model states that ability estimates are obtained on the same scale and that they can be compared although examinees have taken different sets of items from the pool of test items. The second feature of the model states that item statistics do not depend on the sample of examinees used in the calibration of the statistics (Hambleton & Swaminathan, 1985). We investigated whether these two features held for the two-parameter model fitted for the PI-dictee.

First, we addressed the feature of invariance of ability estimates. The PI-dictee was split into two parts, one consisting of the even items and one consisting of the odd items. Two separate models were estimated. The pairs of ability estimates obtained from the two parts of the test for each examinee are plotted in Figure 2. When the item response model fits to the data, the bivariate plot of ability estimates should be linear. From Figure 2, it is clear that such a relationship holds between the ability estimates obtained from the odd and even items. We repeated the same analysis while splitting the test into the known Versions A and B; a similar result was obtained. The conclusion seemed justified that the ability estimates were independent of the chosen items.

Second, we addressed the feature of invariance of item statistics. If the model fits to the data, there should be a linear relationship between the item

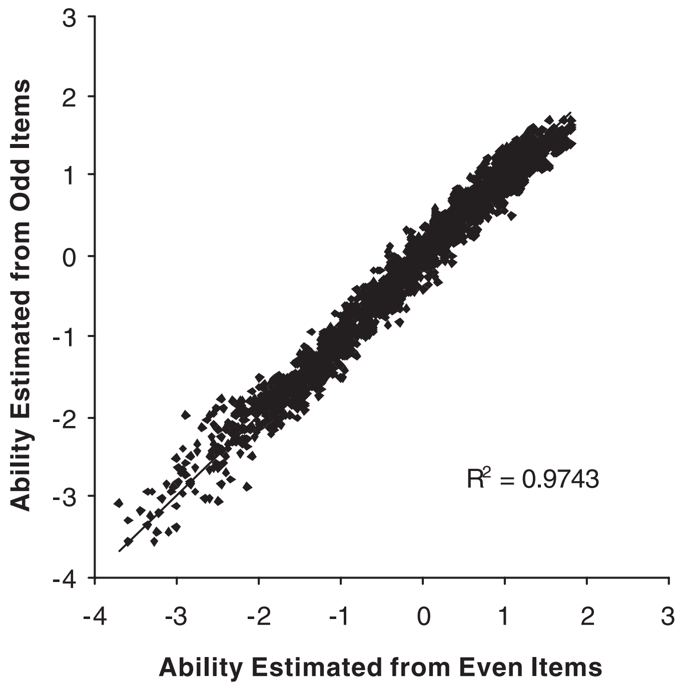

Figure 2. Invariance of ability parameters.

parameters estimated from two samples, even if the samples differ in some respect (Lord & Novick, 1968). We split the group of examinees into boys and girls, and again, we estimated two item response models. The results for the difficulty parameters are given in Figure 3. A linear relationship holds between the estimates obtained from the two different samples.

We performed a separate analysis to investigate whether the two-dimensional structure previously found for the first graders would be reflected in an invariance of the item parameters. We split the items according to the factor loadings found in the analysis of the tetrachoric correlations of the 90-item pairs of Grade 1 analogous to the practice of splitting the items into common content areas (Schatschneider, Francis, Foorman, Fletcher & Mehta, 1999). A total of 71 items loaded mainly on the first factor, and the remaining 19 items loaded on the second factor. Again, the item statistics were estimated twice: once using only the items that were strongly related to the factor of interest and once using all the items. The plots of the difficulty parameters obtained from the separate analyses are given in Figure 4. The highly linear plot indicates that the violation of the assumption of unidimensionality for the first graders did not show a strong effect on the invariance of the difficulty parameters.

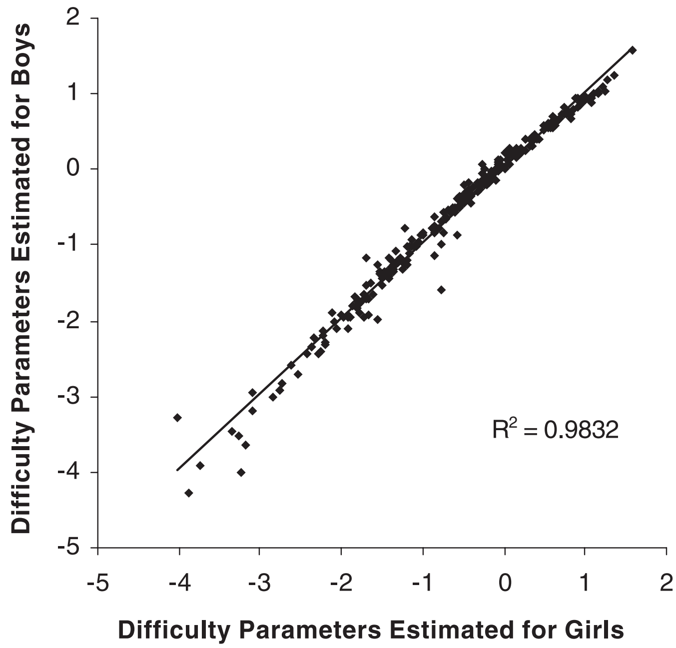

Figure 3. Invariance of difficulty parameters.

Item and Ability Parameters

The average difficulty and discrimination parameters of each block (representing 15 × 2 items of both Versions A and B of the PI-dictee) are presented in Table 5. The average difficulty of the items gradually increases throughout the test. The discriminative power of some items of Block 5 is relatively low; therefore, these items may be removed from the test.

In Table 6, the means and standard deviations on the Ability scale are given for each grade. Remember that during the calibration phase, the means and standard deviation of Grade 1 were arbitrarily set to 0 and 1, respectively. Afterward, the distribution of IRT ability scores was set so that the mean score was 0 and the standard deviation was 1. The data on both the unscaled and scaled ability scores are given in Table 6.

From lower to upper grade levels, the standard deviations diminish and growth in mean scores decelerates. This phenomenon is referred to as scale shrinkage and has, in cases of vertical equating of tests measuring a developmental continuum, been reported before (Camilli, Yamamoto, & Wang, 1993). Several reasons have been discussed for scale shrinkage, including multidimensionality, measurement error, estimation methods, and even IRT

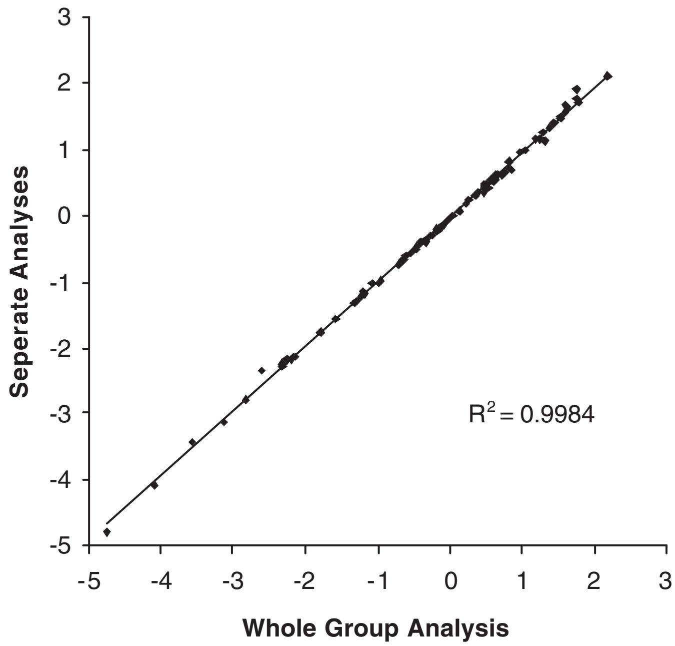

Figure 4. Invariance of difficulty parameters of Grade 1.

modeling itself. Although we do not think that these factors are unimportant here, we considered whether scale shrinkage reflects aspects of growth in skills. It is important that the same pattern in decreasing variances and flattened growth rates is observed for the raw scores (i.e., the number of correctly answered items). In Table 6, the means and coefficients of variations of the raw scores (which express the standard deviation as a percentage of the means) are given for each grade. From this table, it is clear that the relative variances of the raw scores even more sharply decline from the lower to the higher grades than do the relative variances of the IRT ability scores. This suggests that the scale shrinkage is inherent in the developmental continuum studied here.

How well did BILOG-MG perform in vertical equating? This question is important and interesting, but when real testing data are used, it is not easy to answer because no known criteria exist for real data. However, to get an indication of the performance of BILOG-MG, we computed the correlation between the ability scale and the raw score, that is, the number of correctly answered items. The latter measure has the drawback that it may underestimate or overestimate the pupils who are highly superior or inferior achievers, respectively, relative to their grade levels.

Table 5
Average Difficulty and Discrimination Parameters

		Difficulty Parameter		Discrimination Parameter	
Block	Items (Forms A and B)	*M*	*SD*	*M*	*SD*
5	1 to 30	–2.72	0.83	1.75	0.64
10	31 to 60	–1.58	0.34	3.06	0.82
15	61 to 90	–1.25	0.37	3.20	0.85
20	91 to 120	–0.93	0.54	2.83	0.90
25	121 to 150	–0.46	0.44	2.91	0.88
30	151 to 180	–0.03	0.37	2.68	0.65
40	181 to 210	0.24	0.41	3.92	0.73
50	211 to 240	0.37	0.61	2.95	1.21
60	241 to 270	0.85	0.28	3.58	1.12

Table 6
Mean, Standard Deviation, and Relative Variance (RV) of the Ability Scale and Raw Scores

	Ability Scale (scaled)		Ability Scale (unscaled)			Raw Scores		
Grade	*M*	*SD*	*M*	*SD*	RV	*M*	*SD*	RV
1	–1.68	0.52	0	1		46.47	18.12	0.39
2	–0.59	0.41	1.89	0.72	0.38	110.53	19.97	0.18
3	0.09	0.40	3.01	0.67	0.22	166.24	22.44	0.13
4	0.49	0.34	3.72	0.61	0.16	201.65	19.78	0.10
5	0.78	0.31	4.16	0.47	0.11	229.60	20.74	0.09
6	1.01	0.30	4.79	0.64	0.13	242.26	18.26	0.08

The relationship between the raw scores and the IRT ability estimates obtained turned out to be quite strong (r = .98), but the graph shows some strange wings in the upper half. These wings represent the pupils whose IRT ability estimates were higher than were their raw scores. They are the overachievers for whom the items were too easy. For Grades 1 to 4, such a group of overachievers is discernable in Figure 5; for the fifth and sixth graders, no such ceiling effects occur because the most difficult items have been administrated to them. Figure 5 shows an additional remarkable difference between the raw scores and the IRT ability scores. On the lower end of the scale, pupils got, compared with the number of correctly answered items, a relatively lower score on the IRT ability scale. The nonlinear association between the two variables can be explained by the fact that the raw scores are bounded by zero and a maximum score, which leads the IRT estimates to spread out the raw score extremes.

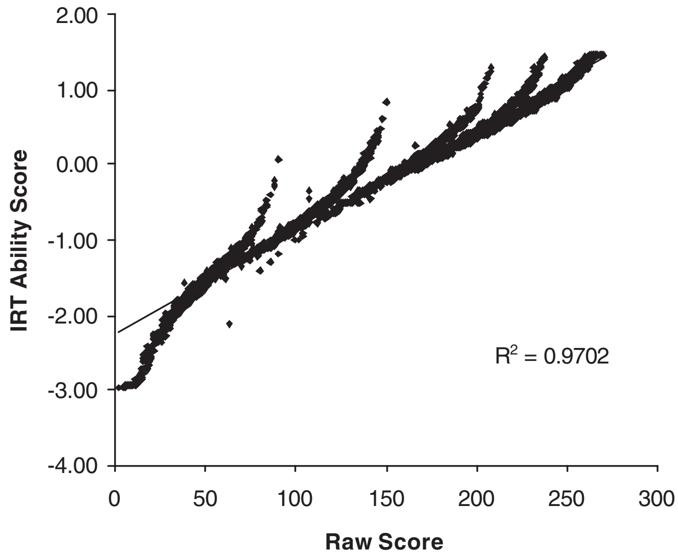

Figure 5. The relationship between item response theory (IRT) ability and raw score.

Information Function

Within IRT modeling, the test information function is quite important as a measure of the precision of the test scores. The amount of information a test provides is influenced by the quality and number of test items. The steeper the slope is (e.g., the greater the *a* parameter), the smaller the item variance is; more items also typically lead to greater information (Hambleton & Swaminathan, 1985). For a given ability level θ, the error associated with ability estimates is inversely related to the amount of information provided by a set of test items.

Both the standard error of the ability estimates at ability level θ and the test information function are given in Figure 6. For the PI-dictee, the amount of information finds its peak at an ability score of zero, which is slightly less than the average score of the third graders. At the extremes of the scales, the information is low and the error of measurement high; nevertheless, the normal range of scores within the elementary school is well covered regarding the amount of information provided.

Within IRT modeling, an equivalent to the reliability coefficient forms the marginal reliability (Thissen, 1990). For the whole test, including all pupils, this coefficient reached .99, which is high.

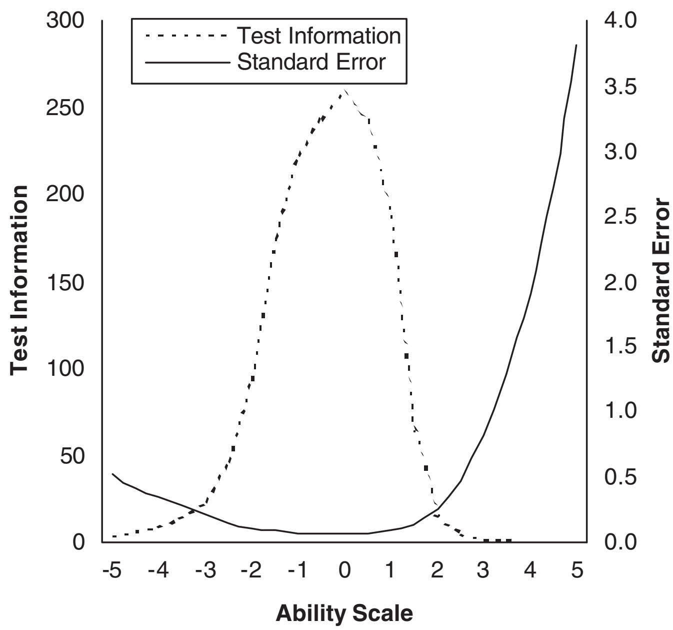

Figure 6. The test information function and standard error.

Discussion

In this study, we investigated whether spelling ability can be conceptualized as a unidimensional construct within the framework of latent trait theory. Factor analyses and IRT modeling were used to address this question. For Grades 2 to 6, the analyses showed that one latent trait is underlying the subsets of items that were administered. However, for the first graders, the solution revealed two factors. Inspection of the factor loadings showed that items heavily loading on the first factor were (almost) all phonologically transparent, whereas items related to the second factor showed some deviance from transparency. This solution was not found for the second graders who completed the same items. Unidimensionality does not depend on the particular items only but is also dependent on the characteristics of the sample. Although a two-factor solution was found for the first graders, the invariance of the item parameters in the subsequent IRT analysis was well preserved.

These results are in line with facets of existing theories of spelling ability development. Besides the knowledge of basic phoneme-grapheme correspondences, spelling involves the application of orthographic knowledge (Lennox & Siegel, 1994). Second graders already implemented orthographic patterns in the spellings of words in a successful way. The division in a phonological and a lexical component previously found for the first graders was absent for the second graders, although the same items were presented to both groups. These results indicate that there is a transition in spelling strategies children use that occurs between the second half of the first grade and the first half of the second grade. First, children spell alphabetically, but fairly soon, they start implementing orthographic patterns in their spellings. The fact that the phonological component was not found for second graders and older students means that because the alphabetic principle was mastered by these pupils, they spelled all phonologically transparent words correctly such that there was no variation on this component anymore.

Although several researchers (Lennox & Siegel, 1994; Treiman 1994) pointed out that children show sensitivity to orthographic patterns from the beginning, our findings probably also reflect some properties of Dutch language and instruction. Because the Dutch language is mainly based on the alphabetic principle, the formal education of reading and spelling starts with that, but at the end of the first grade, a beginning is made with some fairly consistent orthographic patterns and rules. In our results, this general pattern of spelling education was clearly reflected.

We did not find a qualitative shift around the fifth grade when pupils are supposed to start spelling by analogy (Marsh et al., 1980). The words included in the PI-dictee are all commonly known to children of their age; therefore, the use of spelling of unknown words by analogy was not an issue here. However, our results indicate that the storage (and retrieval) of lexical patterns of known words in memory starts at an early point in spelling development.

Although the PI-dictee was two-dimensional for the first graders, we fitted a unidimensional IRT model for all grades of elementary education. In general, IRT modeling was a useful tool in constructing a spelling test that reflected the ability spectrum of the elementary school. The results indicated that the items of the test varied in difficulty as well as in discriminative value. Furthermore, the item and ability estimates proved to be fairly stable in different samples. The information function showed that the amount of information and the standard error are maximized and minimized, respectively, at the ability scale slightly above the average ability of the third graders. The informative value of the test for all grades of elementary school is satisfactory, although the scores for the low-achieving first graders and especially for the high-achieving sixth graders may be less reliable, indicating ceiling effects at the extremes of the test.

Two studies indicated that concurrent calibration as is implemented in BILOG-MG performed equally well in comparison with traditional methods in vertical equating (Béguin & Hanson, 2001; Kim & Cohen, 1998). Because in our study the number of anchor items was large and because these anchor items did form a representative sample of the test, technical issues of vertical equating should not be a problem here.

The results of our study indicate that spelling ability can be conceived as one latent construct, but in the onset of development, it is composed of a phonological and a lexical factor. The importance of the lexical factor in learning Dutch spelling comes rather early in the developmental continuum. Although spelling ability may involve the collection of heterogeneous knowledge elements (e.g., orthographic patterns, word-specific knowledge, the application of spelling rules, and so forth) the results of our study indicate that it can be conceived as one latent construct. In the onset of development, spelling ability is composed of a phonological and lexical factor, whereas the importance of the lexical factor in learning Dutch spelling comes rather early in the developmental continuum. From this study, the conclusion seems to be justified that spelling throughout the elementary grades mainly consists of the implementation of orthographic knowledge.

References

Béguin, A. A., & Hanson, B. A. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating* (Measurement and research department reports). Arnhem, Netherlands: Citogroep. Retrieved September 2002 from http://download.cito.nl/pub/pok/reports/Report01-02.pdf

Bock, R. D., & Zimowski, M. F. (1995). Multiple group IRT. In W. van der Linden & R. Hambleton (Eds.), *Handbook of item response theory* (pp. 433-448). New York/Berlin: Springer-Verlag.

Brown, G. D. A., & Loosemore, R. P. W. (1994). Computational approaches to normal and impaired spelling. In G. D. A. Brown & N. C. Ellis (Eds.), *Handbook of spelling* (pp. 319-335). New York: John Wiley.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, *17*, 379-388.

Caravolas, M., Hulme, C., & Snowling, M. J. (2001). The foundations of spelling ability: Evidence from a 3-year longitudinal study. *Journal of Memory and Language*, *45*, 751-774.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, *25*, 31-45.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, *68*, 363-373.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement*, *5*, 175-186.

Frith, U. (1980). Unexpected spelling problems. In U. Frith (Ed.), *Cognitive processes in spelling* (pp. 495-515). San Diego, CA: Academic Press.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. E. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia* (pp. 301-326). Boston: Routledge Kegan Paul.

Geelhoed, J., & Reitsma, P. (1999). *PI-dictee*. Lisse, the Netherlands: Swets & Zeitlinger.

Guskey, T. R. (1981). Comparison of a Rasch Model scale and the Grade-Equivalent scale for vertical equating of test scores. *Applied Psychological Measurement*, *5*, 187-201.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Dordrecht, the Netherlands: Kluwer Academic.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harris, D. (1991). A comparison of Angoff's Design I and Design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement*, *28*, 221-235.

Henderson, E. H., & Templeton, S. (1986). A developmental perspective of formal spelling instruction through alphabet, pattern and meaning. *Elementary School Journal*, *86*, 305-316.

Holmes, S. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement*, *19*, 139-147.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*, 131-143.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, *18*, 1-11.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York/Berlin: Springer-Verlag.

Lennox, C., & Siegel, L. S. (1994). The role of phonological and orthographic processes in learning to spell. In G. D. A. Brown & N. C. Ellis (Eds.), *Handbook of spelling*. (pp. 93-109). New York: John Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179-193.

Marsh, G., Friedman, M., Welch, V., & Desberg, P. (1980). The development of strategies in spelling. In U. Frith (Ed.), *Cognitive processes in spelling* (pp. 339-353). San Diego, CA: Academic Press.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*, 81-91.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Maximum likelihood item analysis and test scoring with logistic models*. Mooresville, IN: Scientific Software.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: American Council on Education.

Schatschneider, C., Francis, D. J., Foorman, B. R., Fletcher, J. M., & Mehta, P. (1999). The dimensionality of phonological awareness: An application of item response theory. *Journal of Educational Psychology*, *91*, 439-449.

Siegler, R.S. (2000). The rebirth of children's learning. *Child Development*, *71*, 26-35.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, *56*, 495-529.

Templeton, S., & Morris, D. (2000). Spelling. In M. L. Kamil, P. B. Mosenthal, D. P. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 525-543). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 161-185). Hillsdale, NJ: Lawrence Erlbaum.

Treiman, R. (1994). Sources of information used by beginning spellers. In G. D. A. Brown & N. C. Ellis (Eds.), *Handbook of spelling* (pp. 93-109). New York: John Wiley.

Varnhagen, C. K., McCallum, M., & Burstow, M. (1997). Is children's spelling naturally stage-like? *Reading and Writing: An Interdisciplinary Journal*, *9*, 451-481.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International, Inc.